Hacking AI on the Cloud

Lisa Amini, PhD
Lab Director, IBM Research Cambridge
We are in a period of rapid innovation and emerging adoption of data science, machine learning and AI.
Enterprise AI presents unique and challenging requirements
Challenges of Building and deploying AI Models Today

**Training**
- “Roll your own” home-brew environments
- Stateful, compute-intensive execution at odds with cloud-native design
- Stresses cloud networking, storage, and hardware
- Open source components evolving at different rates and speed

**Deployment**
- Testing and debugging neural nets is an active research topic
- Model evolution based on feedback is unavailable
- Enterprise readiness for compliance and traceability is not well understood

**Inference**
- Must handle streaming data
- Near-real-time response required though inferencing on large deep learning networks is compute intensive
- Must be able to run in the cloud and at the edge
Don’t want to:
• Configure machines with GPUs
• Manage the frameworks
• Handle failure and recovery
• Manage load balancing, storage, security and logging
• Manage distributed processing

Data Scientists
Want to:
• Use DL frameworks they love
• Train large models with big data
• Rapidly experiment with different models
• Focus on business logic
• Get the best accuracy

Business Owners
Want to:
• Provide cognitive capabilities to my users
• Rapidly innovate product features
• Access cutting edge AI algorithms
• Be adaptive and flexible
3 approaches for building AI

1. pre-trained AI
   - pre-trained model
   - app developer or SME

2. transfer learning
   - pre-trained model
   - app developer or SME
   - your domain data

3. custom AI
   - data scientist
   - your domain data
   - custom model

Watson Visual Recognition
Natural Language Understanding
Watson Speech to Text
Watson Text to Speech

Watson Visual Recognition
Natural Language Classifier
Watson Speech to Text

Watson Studio
Watson Machine Learning
Deep Learning as a Service

...
Watson Developer Cloud APIs
Pretrained AI

- * Watson Assistant
  - Natural Language Classifier
  - Language Translator
- Personality Insights
- Tone Analyzer
- Natural Language Understanding / AlchemyLanguage
  - Includes Emotion analysis

- * Discovery
  - Discovery News

- Speech to Text
- Text to Speech

- Visual Recognition

* Higher level APIs that include tooling to simplify implementation

For the latest view of Watson API’s available, go to:  www.ibm.com/watsondevelopercloud

https://console.bluemix.net/catalog?category=watson
Deep Learning

Design complex neural networks then experiment at scale to deploy optimized deep learning models, within Watson Studio.

Try Watson Studio for free

What is Deep Learning?

IBM's experiment-centric deep learning service within Watson Studio allows data scientists to visually design their neural networks and scale out their training runs while auto- allocation means paying only for the resources utilized. Optimized for production environments, scale up your training using the NVIDIA® Tesla® V100 GPU with your preferred deep learning framework then easily deploy to the cloud or at the edge.

<table>
<thead>
<tr>
<th>Deep Learning features</th>
<th>Experiment Assistant</th>
<th>Open and flexible</th>
<th>Elastic GPU compute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initiate and monitor batch training experiments then compare cross-model performance in real-time without worrying about log transfers and scripts to visualize results. You focus on designing your neural networks. We'll manage and track your assets.</td>
<td>Use your preferred deep learning framework: TensorFlow, Keras, PyTorch, Caffe and more. Manage your deep learning experiments with the tools you prefer: command-line interface (CLI), Python library or an interactive user interface.</td>
<td>Train neural networks in parallel using market-leading NVIDIA® Tesla® GPUs - K80, P100, and V100. Pay only for what you use. means no more remembering to shutdown your cloud training instances. No clusters or containers to manage.</td>
<td>Let's talk</td>
</tr>
</tbody>
</table>
Create a Watson Studio account: https://www.ibm.com/cloud/watson-studio

Watson Studio
Build, train, deploy and manage AI models, and prepare and analyze data, in a single, integrated environment.

Start your free trial  Watch Deep Learning webinar

What is Watson Studio?
IBM Watson Studio accelerates the machine and deep learning workflows required to infuse AI into your business to drive innovation. It provides a suite of tools for data scientists, application developers and subject matter experts to collaboratively and easily work with data and use that data to build, train and deploy models at scale.
Several ways to build custom AI

“non-coder/clicker” data scientists
- SPSS Modeler
- Neural Network Modeler

“coder” data scientists/researchers
- Jupyter Notebooks
- R Studio

Programming frameworks
- Keras
- TensorFlow
- Caffe
- PyTorch
- Spark
- XGBoost
- PyCharm
- VI
- Terminal
IBM's deep learning solution assists you in scaling to hundreds of training runs while only paying for the resources you require. Designed to fit your existing workflow, select your preferred toolset to design your neural networks then initiate and monitor your training experiments.

- Documentation
- Supported libraries and version

Get Started

**Python Client**

Use Python 3.5 and the Watson Machine Learning library to initiate experiments and monitor your training runs using your preferred coding environment.

[View Python Details]

**Command Line**

Use the command line interface (CLI) to initiate experiments and monitor your training runs.

[View CLI Details]

**Neural Network Modeler**


[Create Network]

**Experiment Builder**

Quickly define training runs for your experiment. Track hyperparameters then automatically optimize them. Monitor training progress and view real-time event graphs.
Deep Learning on the Cloud (DLaaS): Overview

DLaaS

- A “serverless” user experience: focus on Neural Networks and data
- Adaptive deep learning as a service: cloud-native support for popular frameworks; elastic, resilient, and SLA driven
- Enabling quick time-to-value for a fast evolving field by architecting for continuous experimentation.

**User experience and tools**
- Experiment Studio (visualization, multiple parallel jobs, autonn, model governance and management)

**Innovation in neural net design**
- DLaaS supports multiple popular DL frameworks
- Advances in DL frameworks
- Improvement in training techniques
- High performance distribution
- Advances in the cloud stack
- GPU enabled kubernetes (Armada), advanced scheduling

**New hardware**
- New Accelerators

**API (WML): train/manage/watch**
- theano
- TensorFlow
- Caffe
- Keras

**AI Studio/ Data Science Experience**
- Watson Services

**Infrastructure (Softlayer)**
- Armada (GPU enabled)

**Container and resource management**

[Links]
- https://dlaas-api.stage1.mybluemix.net
- https://dlaas-guide.stage1.mybluemix.net
Deep Learning on the Cloud: Goal

Where data scientists run deep learning experiments

Run experiments, provide model tuning and comparison tools to evaluate models across 100-1000’s of hyperparameter configurations
Model training distributed across containers
Multi-tenant container orchestration

containers

Kubernetes container orchestration

server cluster

NVIDIA GPUs

training runs

dataset

Cloud Object Storage
**Key Concepts**

<table>
<thead>
<tr>
<th><strong>Watson Machine Learning Service (WML)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ IBM WML is a comprehensive solution that is your gateway to various machine and deep learning technologies.</td>
</tr>
</tbody>
</table>
| ✓ Provides the following capabilities  
  - Automated model building  
  - Model training and optimization  
  - Model deployment |
| ✓ To interact with WML, you need to create a WML instance |

<table>
<thead>
<tr>
<th><strong>Cloud Object Storage (COS)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Object storage is the primary data resource used in the cloud, and it is also increasingly used for on-premise solutions.</td>
</tr>
<tr>
<td>✓ Inexpensive, scalable, self-healing storage for massive amounts of unstructured data.</td>
</tr>
<tr>
<td>✓ IBM Cloud Object Storage can support single objects as large as 10TB.</td>
</tr>
</tbody>
</table>
Getting Started: Online Docs, Tutorials, Blogs, ...

https://console.bluemix.net/docs/tutorials/index.html#tutorials
Questions?
Connect and discover content from multiple data sources in the cloud or on premises. Bring structured and unstructured data to one toolkit.

Find data (structured, unstructured) and AI assets (e.g., ML/DL models, notebooks, Watson Data Kits) in the Knowledge Catalog with intelligent search and giving the right access to the right users.

Clean and prepare your data with Data Refinery, a tool to create data preparation pipelines visually. Use popular open source libraries to prepare unstructured data.

Democratize the creation of ML and DL models. Design your AI models programmatically or visually with the most popular open source and IBM ML/DL frameworks or leverage transfer learning on pre-trained models using Watson tools to adapt to your business domain. Train at scale on GPUs and distributed compute.

Deploy your models easily and have them scale automatically for online, batch or streaming use cases.

Monitor the performance of the models in production and trigger automatic retraining and redeployment of models. Build Enterprise Trust with Bias Detection, Mitigation Model Robustness and Testing Service Model Security.
Cloud Object Storage

- Object storage is the primary data resource used in the cloud, and it is also increasingly used for on-premise solutions.
- Inexpensive, scalable, self-healing storage for massive amounts of unstructured data.
- IBM® Cloud Object Storage can support single objects as large as 10TB when using multipart uploads.
1. Create model in DL framework supported by DLaaS (Caffe, Tensorflow, Torch, Theano …)

2. Store training data in object storage

3. Specify model metadata (framework, ...); resource requirements (GPUs, mem, num learners, ...); pointer to training data (object store, s3, ...)

4. Start a training job

5. Query training status and retrieve logs

6. Get trained model

DLaaS API: /v1/models

Kubernetes cluster runs training jobs
How to access TensorBoard?

https://www.tensorflow.org/programmers_guide/summaries_and_tensorboard

To access TensorBoard, you can use the following steps:

1. Download the training run files:
   ```
   python download_training_run_files.py TRAINING_RUN
   ```
2. Start TensorBoard with the log directory:
   ```
   tensorboard --logdir=path/to/log-directory
   ```
3. View TensorBoard at:
   ```
   http://localhost:6006
   ```
<table>
<thead>
<tr>
<th>AI algorithms</th>
<th>Physics of AI</th>
<th>Applications of AI to industries</th>
<th>Advancing shared prosperity through AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning and Reasoning</td>
<td>Analog AI</td>
<td>Healthcare, Life Sciences</td>
<td>Ethics of AI, fair, unbiased</td>
</tr>
<tr>
<td>Continuous, multi-task, small data, explanations, ...</td>
<td>AI &amp; Quantum</td>
<td>Cybersecurity</td>
<td>AI for Social Good</td>
</tr>
</tbody>
</table>

**MIT-IBM Watson AI Lab**

$240M 10 year commitment to jointly create the future of artificial intelligence

[http://mitibmwatsonailab.mit.edu/](http://mitibmwatsonailab.mit.edu/)